Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2019, 7, 174-189
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Hot Events Detection of Stock Market Based on Time Series Data of Stock and Text Data of Network Public Opinion

Beibei Cao

Department of Publishing and Dissemination, Shanghai Publishing and Printing College, Shanghai, China
Email: bbc@sppc.edu.cn

How to cite this paper: Cao, B.B. (2019) Hot Events Detection of Stock Market Based on Time Series Data of Stock and Text Data of Network Public Opinion. *Journal of Data Analysis and Information Processing*, **7**, 174-189.
https://doi.org/10.4236/jdaip.2019.74011

Received: July 11, 2019
Accepted: September 27, 2019
Published: September 30, 2019

Abstract

With the highly integration of the Internet world and the real world, Internet information not only provides real-time and effective data for financial investors, but also helps them understand market dynamics, and enables investors to quickly identify relevant financial events that may lead to stock market volatility. However, in the research of event detection in the financial field, many studies are focused on micro-blog, news and other network text information. Few scholars have studied the characteristics of financial time series data. Considering that in the financial field, the occurrence of an event often affects both the online public opinion space and the real transaction space, so this paper proposes a multi-source heterogeneous information detection method based on stock transaction time series data and online public opinion text data to detect hot events in the stock market. This method uses outlier detection algorithm to extract the time of hot events in stock market based on multi-member fusion. And according to the weight calculation formula of the feature item proposed in this paper, this method calculates the keyword weight of network public opinion information to obtain the core content of hot events in the stock market. Finally, accurate detection of stock market hot events is achieved.

Keywords

Relationship, Network Public Opinion, Stock Trading Behavior, Stock Market Hot Events

1. Introduction

In the securities industry, once market fluctuations occur, investors first hope to find the answer from the Internet information. However, the geometric expan-

sion of Internet information makes it more and more difficult for people to extract effective information. If investors are unable to obtain timely and accurate information about events that lead to financial market volatility, then the losses caused are incalculable. Therefore, how to quickly find valuable topics and events from a large number of Internet data is particularly important.

With time goes by, numerous research methods for event discovery have been put forward [1]-[11]. However, most of these methods are based on text data [1]-[11] or time series data [12]-[19] for event discovery separately. There are few scholars, to the best of our knowledge, study the characteristics of financial time series data and text data to carry out research [20] [21] [22]. As a realistic behavior of financial markets, time-series data such as stock trading data and market data are often affected by events and can better reflect changes before and after events. Therefore, this paper studies the discovery of financial events by combining network text information and financial time series data, so as to help investors to quickly obtain hot events and correctly grasp market dynamics.

2. Post's Influence of Network Public Opinion Space

2.1. Definition and Quantification of Post's Activity

In web forums, netizens can express their concern for specific information by posting, reading and replying. And this degree of attention is an important external feature of the emotional tendency of network public opinion. In this paper, we call it post's activity. In order to quantify the user's attention to topic information intuitively, we calculate it by the amount of readings and the amount of comments of the posts. Among them, the readings amount of posts reflects the degree of dissemination of the information contained in the posts and it is the instinct concern of users. The comments amount of posts reflects the attention paid to the information contained in the posts. And it is the manifestation of the user's emphasis on topic interaction, and its emotional intensity is stronger. So in this paper, we choose the amount of readings and the amount of comments as indicators of post's activity. The specific definitions are as follows:

Definition 2-1 Post's activity: Assuming that within a period of time t, a total of N posts are posted in the online public opinion space, which are $\{p_1, p_2, \cdots, p_i, \cdots, p_n\}$. The readings amount of the i-th post is p_i_r, and the comments amount is p_i_c. Then the total readings amount of N posts in the time period t is $r = \sum_{i=1}^{N} p_i\_r$, and the average readings amount of each post is $avg\_r = \frac{r}{N}$. We define the propagation coefficient p_i_p of p_i as the ratio of the readings amount of p_i in time period t to the average reading amount of each post in the same time period. The formula of p_i_p is $p_i\_p = \frac{p_i\_r}{avg\_r}$. Similarly, the attention coefficient p_i_c of p_i is defined as the ratio of the comment amount of p_i in time period t to the average comment amount of

each post in the same time period, and the formula is $p_i\_c = \frac{p_i\_c}{avg\_c}$, where $avg\_c = \frac{c}{N} = \frac{\sum_{i=1}^{N} p_i\_c}{N}$. Finally, the activity of p_i is defined as the sum of its propagation coefficient and the attention coefficient, namely:

$$\begin{aligned} p_i\_Ac &= p_i\_p + p_i\_c \\ &= \frac{p_i\_r}{avg\_r} + \frac{p_i\_c}{avg\_c} \\ &= \frac{p_i\_r}{\frac{\sum_{i=1}^{N} p_i\_r}{N}} + \frac{p_i\_c}{\frac{\sum_{i=1}^{N} p_i\_c}{N}} \end{aligned} \tag{1}$$

2.2. Definition and Quantification of User's Influence

The influence of users in the stock bar forum refers to the popularity index of the user in the stock bar. It is mainly affected by the age of the user, the amount of comments posted by the user, the amount of forwarding, and other factors. So in this paper we use user's power, user's activity and user's attention to measure user's influence in the stock bar forum.

2.2.1. User's Power

User's power is the potential energy that users have under static conditions. It is mainly reflected in the three factors of age, the amount of fans and the amount of people that user concern.

Definition 2-2 User's power: The user's power of the *i*-th user a_i is defined as $a_i = \frac{\frac{Pa_i}{\overline{Pa}} + \frac{Pfr_i}{\overline{Pfr}} + \frac{Pfe_i}{\overline{Pfe}}}{3}$, where Pa_i is the age of the user a_i, $\overline{Pa}$ is the average age of all users, Pfr_i is the amount of fans of the user a_i, $\overline{Pfr}$ is the average amount of fans of all users, Pfe_i is the amount of people that user a_i concern, and $\overline{Pfe}$ is the average amount of people that all users concern.

2.2.2. User's Activity

User's activity reflects the degree of user's autonomy, which is mainly determined by the amount of postings and the amount of comments. A new post usually contains a new topic. Therefore, the more posts a user publishes, the easier it is for other users to pay attention to the post, and the greater the influence of the user. Comments reflect the user's views and opinions about other people's information, and it is also a manifestation of user's activity.

Definition 2-3 User's activity: The user's activity of the *i*-th user a_i is defined as $a_i\_Ac = \frac{\frac{Pp_i}{\overline{Pp}} + \frac{Pd_i}{\overline{Pd}}}{2}$, where Pp_i represents the total number of posts for the user a_i, $\overline{Pp}$ represents the average number of posts for all users, Pd_i represents the total number of comments for the user, and $\overline{Pd}$ represents the

average number of comments for all users.

2.2.3. User's Attention

User's attention mainly reflects the degree of attraction and attention of users to other users in the online forum. When a user's posts are commented by a large number of other users, it shows that the quality of these posts are high and attractive, which further indicates that the user has great influence. In addition, there are some users who are not good at commenting, but are used to expressing their concern about posts through reading, which also shows the attraction of posts to them. Therefore, the total amount of readings also needs to be regarded as the influencing factor of users' attention.

Definition 2-4 User's attention: The user's attention of the i-th user a_i is defined as $a_i\_At = \frac{\frac{Pr_i}{\overline{Pr}} + \frac{Pc_i}{\overline{Pc}}}{2}$, where Pr_i represents the average readings amount of all posts of user a_i, $\overline{Pr}$ represents the average readings amount of all users' posts, Pc_i represents the average comments amount of all posts of user a_i, and $\overline{Pc}$ represents the average comments amount of all users' posts.

Based on the above three user indicators, the paper calculates the user's influence formula as follows:

$$a_i\_I = a_i\_P + a_i\_Ac + a_i\_At$$

In the end, we calculate the post's influence of network public opinion space by combining the activity of the post and the influence of the poster. The calculation formula is as follows:

$$p_i\_I = 0.7 \times p_i\_Ac + 0.3 \times a_i\_I \quad (2)$$

3. Hot Event Detection of Stock Market

3.1. Definition of Stock Market Hot Events

The events to be studied in this paper refer to hot events that are related to the stock market and can lead to changes in stock trading behavior. This paper defines it as hot events of stock market. It is embodied in the following three characteristics:

1) The events corresponding to popular posts (which have been read and commented for many times and have high influence) on the web forums.

2) The events corresponding to online hot news (which is reported or reproduced by multiple news websites).

3) The events that can have a significant impact on the stock market.

The first two are based on the feedback from the online public opinion space to understand the stock market hot events, and the third is to understand the hot events of stock market based on the information fed back from the real trading conditions of the stock market. These three event characteristics will be fully combined below, and on this basis, we will conduct a research on the detection of hot events in the stock market.

3.2. Time Extraction of Stock Market Hot Events Based on Multi-Member Fusion

The trajectory of events in the real world is often reflected by the trajectory of the events-related members. Therefore, whether the abnormal changes in the state of the event members can be found is the key to determine whether or not the event occurs. According to the definition of stock market hot events in Section 3.1, this paper studies the relationship between stock trading behavior attributes, post's influence and online news volume as relevant members of stock market hot events. And based on this, an event time series can be constructed. Among them, the specific relationship between the attributes of stock trading behavior is obtained by the previous research of our team [23] [24]. The research found that there are a certain relationship among the four pairs of attributes in stock attributes, which are Vol and Close, Pcl and Open, Close and %Tuv, %Tuv and %Chg. By detecting the abnormal points of the event time series, we can discover the occurrence time of hot events in the stock market. The event time series is defined as follows:

Definition 3-1 Event Time Series (ETS): Given an event time series $E=\{\langle T_1,M_1\rangle,\langle T_2,M_2\rangle,\cdots,\langle T_i,M_i\rangle,\cdots,\langle T_n,M_n\rangle\}$ where T represents the time at which the event occurred, $M=\{m_1,m_2,\cdots,m_k\}$ represents the set of members of the event, which consists of a total of k event members, n represents the length of the event time series.

In this paper, the event member set consists of six event members, which are the relationship between Vol and $Close$, the relationship between Pcl and $Open$, the relationship between $Close$ and $\%Tuv$, the relationship between $\%Tuv$ and $\%Chg$, post's influence and online news volume. We use the k-nearest neighbor local anomaly detection algorithm to detect anomaly points in event time series, where each $\langle T_i,M_i\rangle$ can be regarded as a feature point in the data point set D. Then, the local anomaly coefficients of each feature point are calculated, and are sorted according to the value of LOF. According to the ranking results, the feature points with the largest values of the first λ are output, and these feature points constitute a set of abnormal points of the event time series. In this paper, these abnormal points are considered as the occurrence points of hot events in the stock market, and the time corresponding to these points is the occurrence time of the hot events. Figure 1 shows the abnormal point detection process for stock market hot event time series.

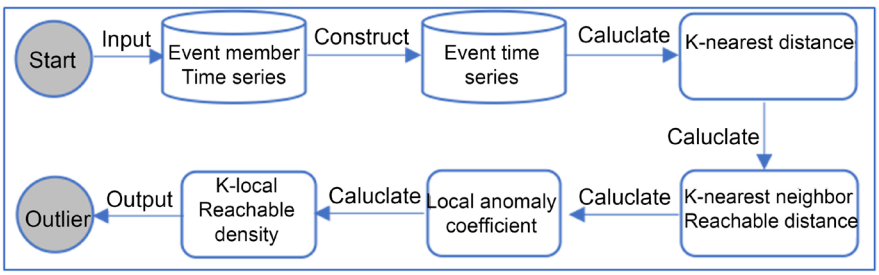

Figure 1. Stock market hot event time series abnormal point detection process.

3.3. Content Extraction of Stock Market Hot Events Based on Multi-Feature Fusion

3.3.1. Weight Calculation of Feature Items Based on Multi-Feature Fusion

In this paper, we choose Vector Space Model (VSM) to vectorize the pre-processed online public opinion text (posts and news). The specific expression is as follows:

$$\begin{array}{c|cccccc} & t_1 & t_2 & \cdots & t_j & \cdots & t_m \\ d_1 & w_{11} & w_{12} & \cdots & w_{1j} & \cdots & w_{1m} \\ d_2 & w_{21} & w_{22} & \cdots & w_{2j} & \cdots & w_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_i & w_{i1} & w_{i2} & \cdots & w_{ij} & \cdots & w_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_n & w_{n1} & w_{n2} & \cdots & w_{nj} & \cdots & w_{nm} \end{array}$$

Each row represents a document, and each column represents a feature item. For each document, it consists of several feature items, which can be represented by a vector $V(d_i)=\{(t_1,w_{i1}),(t_2,w_{i2}),\cdots,(t_j,w_{ij}),\cdots,(t_m,w_{im})\}$. In the expression, d_i denotes the i-th document, t_j denotes the j-th feature item in the document, w_{ij} denotes the weight of the j-th feature item in the document d_i, n denotes the total number of documents, and m denotes the total number of feature items.

The weights of feature items represent the measurement of document content, which is usually calculated by the classical Term Frequency-Inverse Document Frequency (TF-IDF) method [25]. Considering that the feature item set in this paper is mainly obtained by word segmentation of post content and network news content, and each keyword in the set includes valid information such as TF, IDF, and part of speech. Therefore, in this paper, we optimize the original feature weight calculation formula, and calculate the weight of feature items according to the optimized formula. The calculation formula is as follows:

$$\begin{aligned} Weight(d_i,t_j) &= 0.8\times tf(d_i,t_i)\times \lg\left(\frac{D}{n(t_j)}+0.01\right) \\ &\quad +0.2\times Position(d_i,t_j)\times\frac{length(t_j)}{AvgLen} \end{aligned} \tag{3}$$

$$Position(t_i)=\begin{cases}1.5 & \text{when } t_i \text{ is located in the title}\\ 1 & \text{when } t_i \text{ is located in the body}\end{cases}$$

where $Weight(d_i,t_j)$ represents the weight of the feature item t_j in the document d_i. $tf(d_i,t_i)$ represents the frequency of the feature item t_j in the document d_i. $\lg\left(\frac{D}{n(t_j)}+0.01\right)$ represents the inverse document frequency of the feature item t_j. D represents the total number of all documents in the document set. $n(t_j)$ represents the number of documents in a document set that contain feature item t_j. When a feature item appears in each document of the document set, the inverse document frequency of this feature item is 0. In order

to avoid dividing by 0, this paper adds a smoothing factor of 0.01 on this basis. $Position\left(d_i, t_j\right)$ represents the position coefficient of the feature item t_j in the document d_i. Usually an article is composed of two parts, the title and the body. The title is more representative than the body, and it is the overall summary of the article. And when netizens read articles, they usually browse the title first. Therefore, when calculating the weights of feature items, it is necessary to take into account the location factor of the feature item in the document. $length\left(t_j\right)$ represents the length of the feature item t_j, which reflects the amount of information contained in the feature item to a certain extent. In general, the longer the length of a feature item, the greater the amount of information it contains. $AvgLen$ represents the average length of all feature items in the document d_i.

3.3.2. Content Extraction of Stock Market Hot Events

When a hot event occurs in the stock market, the relevant reports and comments on this hot event will grow explosively and update rapidly every day. If the relevant information in a long period of time is selected for event content discovery, the large scale of data and the fast updating of content will result in inaccurate discovery of the event content. In order to discover the stock market hot events accurately, this paper groups the text information related to the events in chronological order, thus constructing a sequence of document collections sorted by time. Each document set in the sequence consists of text documents related to the event in the corresponding time, and the contents of these documents contain all the information of the hot events in that time period. Considering that there may be duplicate or invalid contents in some documents related to the same event topic in the document set, it is necessary to extract key information from each document set of the document set sequence. In this paper, we believe that the key information of the event is mainly composed of several words that have the most contribution to the event information, the most relevant content, and the most abundant information content. Therefore, this paper uses the weight calculation formula proposed in Section 3.3.1 to calculate the weight of vocabulary in each document set, so as to get a set of keywords most relevant to the event content to represent the core content of the event. The method for extracting the hot event content of the stock market is described as follows:

1) Sort the crawled text documents in chronological order and divide them by day, thereby obtaining a sequence of document collections:
$D = \left\{D_1, D_2, \cdots, D_i, \cdots, D_n\right\}$. n represents the time span.
$D_i = \left\{d_1, d_2, \cdots, d_j, \cdots, d_m\right\}$ represents that the document collection on the i-th day consists of a total of m documents.

2) Use the weight calculation formula to calculate the keyword weight of each document d_j in each document collection D_i, and the weight values of the same keywords in the same document collection D_i are accumulated to obtain $V\left(D_i\right) = \left\{\left(t_1, w_{i1}\right), \left(t_2, w_{i2}\right), \cdots, \left(t_j, w_{ij}\right), \cdots, \left(t_m, w_{im}\right)\right\}$. In this expression, $\left(t_j, w_{ij}\right)$ denotes the weight value w_{ij} of the j-th keyword in the document set D_i.

3) According to the weight value, the keywords in each document set are

sorted in descending order, so that the first k keywords of each document set are obtained. These sets contain the core content of the event.

3.4. Detection of Hot Events in Stock Market

In the financial field, time series data and text data are often related and interacted with each other. The emergence of a hot event often leads to changes in stock market and further promotes public opinion. Therefore, when conducting event discovery research for the financial field, we should take into account both the text features and the unique temporal characteristics of the financial field. This paper takes the post information and news information as the breakthrough point, and combines the characteristics of time series data in the stock market to realize the hot spot event detection in the stock market. The overall flow chart for event detection is shown in Figure 2 below. Firstly, construct the event time series by the relationship between stock attributes, post's influence and network news volume. On this basis, the outlier detection algorithm is used to detect the time attributes of the event. Then, the keyword weight of the network public opinion information is calculated by using the feature weight formula of multi-feature fusion to get the event core content of each document set in the document set sequence. Finally, according to the time attributes of events, the corresponding event set is obtained to realize hot event detection in the stock market.

4. Experiments and Results

4.1. Collection and Preprocessing of the Information of Network Public Opinion Space

In this paper, we used the web crawler program to crawl the online post data and the online news data of 21 stocks in the liquor sector from the Eastern Fortune website and Sina Finance website respectively. The details of the post include stock code, title, content, author, time, the amount of readings, author's age, the amount of author's fan, the amount of people that author concerned, the amount of posts posted by the author, the amount of authors comments, the total amount of comments for all posts of the author, and the total amount of readings for all posts of the author. The details of the news include headlines, content, author and time. In this paper, we used ICTCLAS, a Chinese word segmentation system provided by the Chinese Academy of Sciences, to segment the online post data and the online news data.

4.2. Computation and Analysis of the Post's Influence in Network Public Opinion Space

According to formula (1), (2), (3), we calculated all the posts of each stock in turn, so as to get the post's activity, user's influence and the final influence of each post, and then added up the influence of all the posts in each day to get the ultimate influence of posts on that day. As shown in Figure 3, the daily post's influence of 12 stocks over a period of time is shown. From the figure, we can see

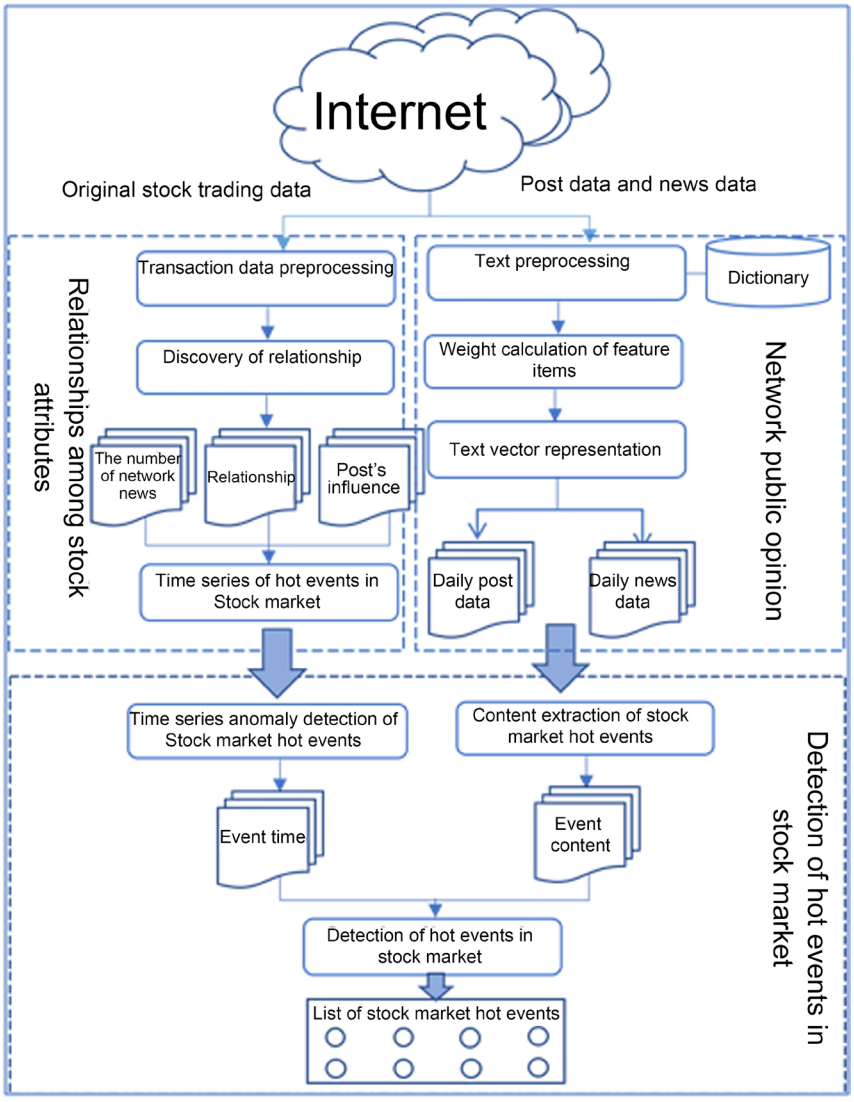

Figure 2. The overall flow chart of hot event detection in stock market.

that the influence of posts is different every day. But most of the time, the difference of the post's influence is not big, and the trend is stable and the value is low. Only in a few time periods will the influence suddenly increase and the value becomes larger. We think this is mainly related to the activities of the stock market. When the stock market is running normally and there is no major change, the netizens only pay attention to and discuss the stock information according to their daily habits, so the influence of related posts will not change much. However, once a major event occurs in the market, it will attract the attention of interested netizens quickly in a short time. At this time, the related posts will be published, commented and forwarded in large quantities, which will lead to a sudden increase in the influence of the post on this day, that is, abnormal. In order to verify this notion, this paper selected 10 historical events of 3 stocks randomly for time comparison. The event table is shown in **Table 1**. **Figure 4** below shows the trend of post's influence of three stocks in the corresponding

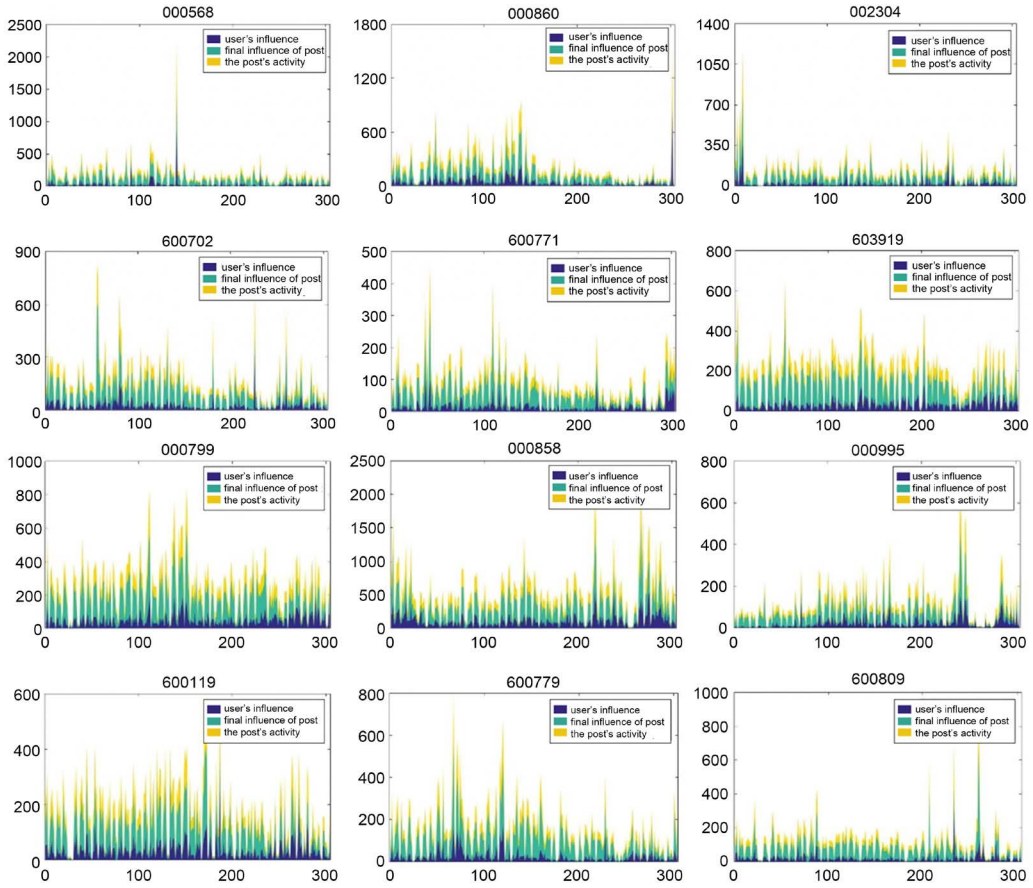

Figure 3. The graph of the daily post's influence of 12 stocks.

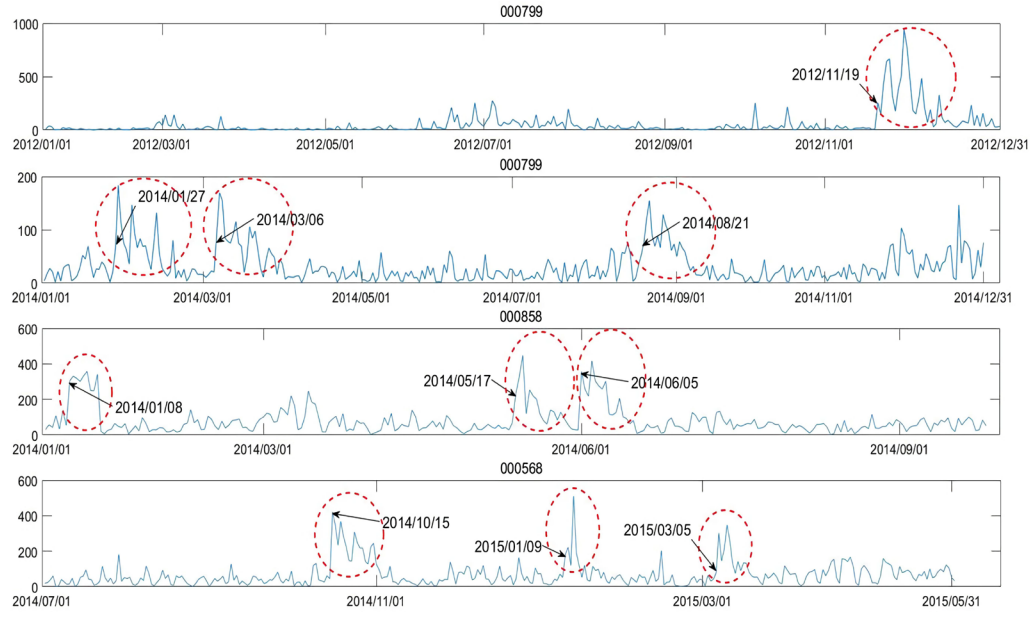

Figure 4. The comparison of the post's influence and the happen time of events.

period of time. In the graph, we mark the abnormal region of the influence of the post with red dotted line, and mark the starting point of the above 10 historical

Table 1. The table of some major historical events from 2012 to 2015.

Stock Code	Event Time	Event Content
000799	2012/11/19	Plasticizer event
	2014/01/27	Alcoholic liquor company's nearly 100 million Yuan deposit was stolen
	2014/03/06	Alcoholic liquor re-generates plasticizer event
	2014/08/21	COFCO Group enters the alcoholic liquor
000858	2014/01/08	Wuliangye Group auction Bus
	2014/05/17	The price of Wuliangye's core products has been lowered
	2014/06/05	Jingdong and Wuliangye reached a strategic cooperative
	2014/10/15	Luzhou Laojiao's bank deposit of 150 million Yuan is abnormal
000568	2015/01/09	Luzhou Laojiao reappears the black swan event—350 million Yuan lost in deposits
	2015/03/05	Directors, financial directors and other three executives resigned

events in the graph. From the graph, we can find that the starting point of historical events coincides with the abnormal points of post's influence. This shows that the influence of the post in this paper can reflect the abnormal changes of stock market to a certain extent, and it is the verification of the above speculation.

4.3. Detection of Hot Events

According to the formula of feature weight introduced in Section 3.3.1, the keywords in each document are calculated sequentially, and the weight values of the same keywords in each document on the same day are added up in time units. Then the keywords are sorted in descending order according to the weight value. Table 2 below shows the top 10 keywords and their weight results of 600519 stock on 2018/05/08. Based on the results, we can conclude that the core event corresponding to the day of 2018/05/08 is "Guizhou Maotai group changes coach. Yuan Renguo resigns and Li Baofang takes office".

After obtaining the daily keyword set of 21 stocks, we need to detect the occurrence time of the hot event. We constructed the event time series according to the introduction in Section 3.2. Some sample data of event time series corresponding to 600519 stock is shown below.

E = {<2018-01-01, 0.8812, 0.8760, 0.8405, 0.8342, 1.8077, 17>,
<2018-01-02, 0.8856, 0.8229, 0.8466, 0.8571, 6.7194, 36>,
<2018-01-03, 0.8447, 0.8738, 0.9120, 0.8503, 2.7293, 21>,
…}

According to the anomaly detection algorithm based on multi-member fusion introduced in Section 3.2, we detected the anomaly points in the event time series, and then calculate the daily LOF value. Figure 5 shows the daily LOF values of 600519 stock over the period from 2018/01/01 to 2018/10/31. As can be seen from the figure, when the abnormal point threshold of the hot event is set to 110, we can get 6 abnormal points, and the corresponding time is 2018/01/15, 2018/05/08, 2018/06/06, 2018/06/24, 2018/09/07, 2018/10/11. In the figure, we

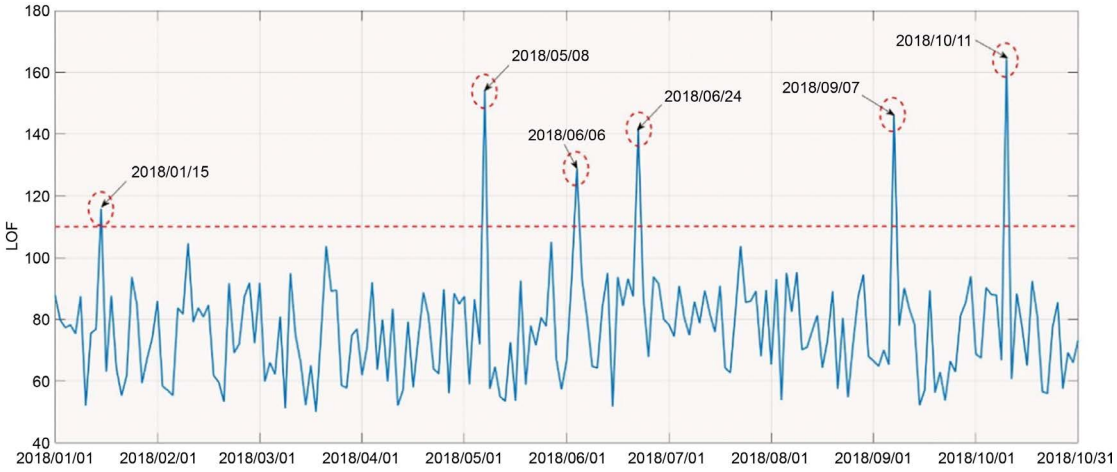

Figure 5. Results of LOF values for 600519 stock in 2018 year.

Table 2. The top 10 keywords and their weight results of 600519 stock on 2018/05/08.

Ranking	Keyword	Documents' number	Total number of documents with keywords	Weight value
1	Maotai	364	217	6.84241
2	change coach	364	195	5.67936
3	Guizhou	364	162	5.37199
4	chairman	364	102	4.68227
5	Li Baofang	364	95	4.30837
6	group	364	88	4.22414
7	take office	364	79	3.39573
8	Yuan Renguo	364	53	2.55493
9	resign	364	47	2.30940
10	pick up	364	23	1.21843

use red dashed lines and text to mark them. We believe that the time corresponding to these six abnormal points is the occurrence time of the stock market hot event of stock 600519.

Based on this result and the daily keyword set obtained above, hot events in the stock market can be found. Table 3 shows the top 5 keywords of each day at the six abnormal time points mentioned above and the content of hot events in the stock market integrated according to these keywords.

The above is the whole process of stock market hot event discovery for 600519 stock by combining the time series data of stock trading and online public opinion text data. Use this process to find stock market hot events on the remaining 20 stocks. Figure 6 shows a graph of the calculation of LOF values for four stocks. And Figure 7 shows the daily LOF values of 21 stocks over the period from 2018/01/01 to 2018/10/31. According to this result, we set the threshold of abnormal points of hot events to 138 to obtain the top ten stock market hot events of the liquor sector stocks in 2018, which are marked with text arrows in the graph. The specific hot event set is shown in Table 4 below.

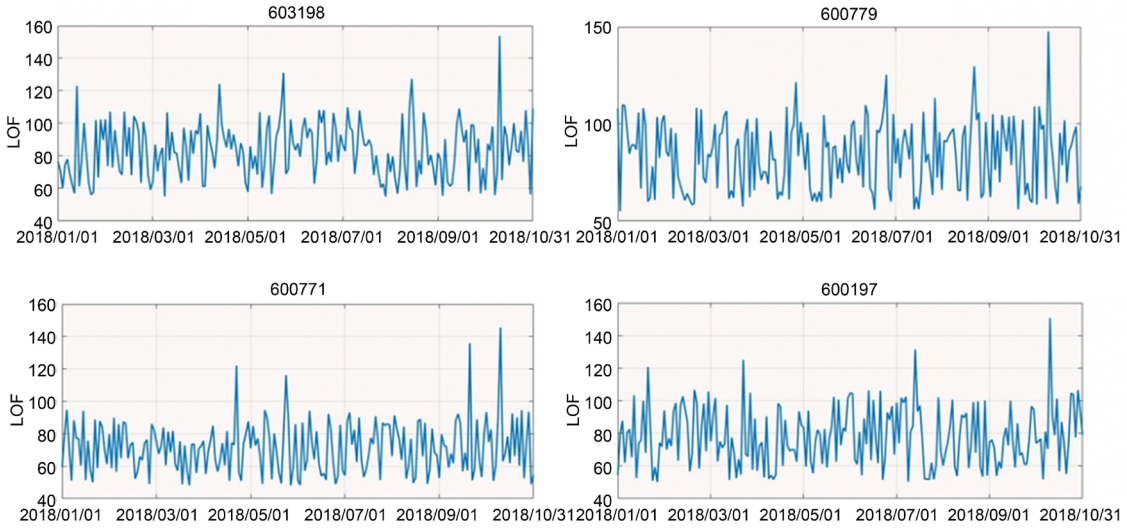

Figure 6. The LOF value of four stocks.

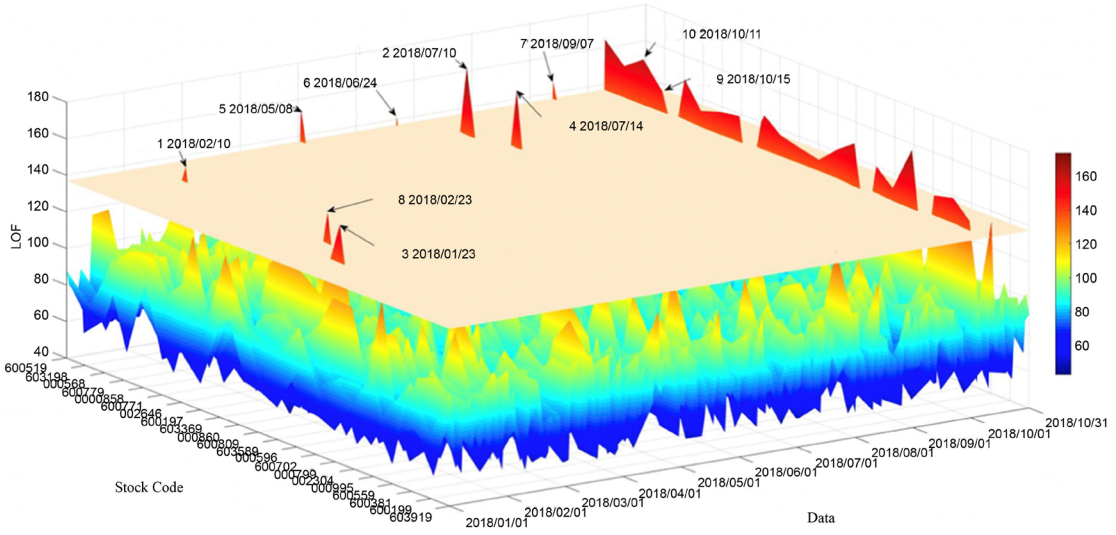

Figure 7. The LOF value of 21 stocks in the liquor sector in 2018.

Table 3. Hot events in the period from 2018/01/01 to 2018/10/31 (600519).

Date	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5	Event content
2018/01/15	Maotai	Market value	Breakthrough	Trillions	High position	Maotai's market value breaks through trillion
2018/05/08	Maotai	Change coach	Guizhou	Chairman	Li Bao Fang	Guizhou Maotai changes coach and Li Baofang become the chairman
2018/06/06	Guizhou	Maotai	Trillions	Market value	Breakthrough	Guizhou Maotai's market value breakthrough trillion again
2018/06/24	Maotai	Bribery	Tan Dinghua	Former vice president	Violation of discipline	Former Vice-President of Maotai Violated Discipline and Accepted Bribes
2018/09/07	Ma Yun	Visit	Maotai	Lead the team	Cooperation	Ma Yun lead the team to visit Maotai at night
2018/10/11	Liquor	Tax increase	Tobacco and alcohol	Rumor	Slump	Rumors of Liquor Tax Increase

Table 4. The top 10 stock market hot events of the liquor sector stocks in 2018.

Number	Stock	Event time	Event content
1	000568	2018/02/10	Luzhou Laojiao acquired a 30% stake in Sichuan Wine
2		2018/07/10	National cellar 1573 ceases to supply
3	000596	2018/01/23	Relapse into the "blending gate" incident
4	000858	2018/07/14	Zhang Hui, director of Wuliangye, was investigated for violation of discipline
5	600519	2018/05/08	Guizhou Maotai changed coach and Li Baofang became chairman
6		2018/06/24	Former vice-president of Maotai violated discipline and accepted bribes
7		2018/09/07	Ma Yun lead his team to visit Maotai in the night
8	600809	2018/02/03	Huarun spent 5.16 billion Yuan to buy a 11.45% stake in Shanxi Liquor
9	603369	2018/10/15	Jinshiyuan acquired a 49% stake in Jingzhi Wine Industry
10	all	2018/10/11	Rumors of tax increase on high-end liquor

5. Conclusion and Future Work

In this paper, we consider that the impact of an event in the financial field will often be mapped to the network public opinion space and the real transaction space at the same time. Therefore, this paper proposes a multi-source heterogeneous information detection method combining stock transaction time series data and network public opinion text data to discover stock market hot events. However, the characteristics of the time series data and text data considered in this paper are still limited. So, in subsequent studies, we can consider combining more temporal and textual features to assist in the discovery of hot events in the stock market.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Lefever, E. and Hoste, V. (2016) A Classification-Based Approach to Economic Event Detection in Dutch News Text. In: *Tenth International Conference on Language Resources and Evaluation*, European Language Resources Association, Paris, 330-335.

[2] Edouard, A. (2017) Event Detection and Analysis on Short Text Messages. Université Côte D'Azur.

[3] Das, S.R. and Chen, M.Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, **53**, 1375-1388. https://doi.org/10.1287/mnsc.1070.0704

[4] Sun, X., Wu, Y., Liu, L., *et al.* (2015) Efficient Event Detection in Social Media Data

Streams. *IEEE International Conference on Computer and Information Technology*; *Ubiquitous Computing and Communications*; *Dependable*, *Autonomic and Secure Computing*; *Pervasive Intelligence and Computing*, Liverpool, 26-28 October 2015, 1711-1717. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.258

[5] Zhao, Y., Qin, B., Liu, T., *et al.* (2016) Social Sentiment Sensor: A Visualization System for Topic Detection and Topic Sentiment Analysis on Microblog. *Multimedia Tools and Applications*, **75**, 8843-8860. https://doi.org/10.1007/s11042-014-2184-y

[6] Atefeh, F. and Khreich, W. (2015) A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, **31**, 132-164. https://doi.org/10.1111/coin.12017

[7] Shi, L., Liu, L., Wu, Y., *et al.* (2017) Event Detection and User Interest Discovering in Social Media Data Streams. *IEEE Access*, **5**, 20953-20964. https://doi.org/10.1109/ACCESS.2017.2675839

[8] Cordeiro, M. (2012) Twitter Event Detection: Combining Wavelet Analysis and Topic Inference Summarization. *Doctoral Symposium on Informatics Engineering*, 11-16.

[9] Nguyen, D.T. and Jung, J.E. (2017) Real-Time Event Detection for Online Behavioral Analysis of Big Social Data. *Future Generation Computer Systems*, **66**, 137-145. https://doi.org/10.1016/j.future.2016.04.012

[10] Li, R., Lei, K.H., Khadiwala, R., *et al.* (2012) Tedas: A Twitter-Based Event Detection and Analysis System. *IEEE* 28*th International Conference on Data Engineering*, Arlington, 1-5 April 2012, 1273-1276. https://doi.org/10.1109/ICDE.2012.125

[11] Phuvipadawat, S. and Murata, T. (2010) Breaking News Detection and Tracking in Twitter. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3, 120-123. https://doi.org/10.1109/WI-IAT.2010.205

[12] Siddique, B. and Akhtar, N. (2017) Temporal Hierarchical Event Detection of Time-Stamped Data. *International Conference on Computing, Communication and Automation*, 5-6 May 2017, 783-788. https://doi.org/10.1109/CCAA.2017.8229902

[13] Jin, F., Wang, W., Chakraborty, P., *et al.* (2017) Tracking Multiple Social Media for Stock Market Event Prediction. In: *Industrial Conference on Data Mining*, Springer, Cham, 16-30. https://doi.org/10.1007/978-3-319-62701-4_2

[14] Alkhamees, N. and Fasli, M. (2017) Event Detection from Time-Series Streams Using Directional Change and Dynamic Thresholds. *IEEE International Conference on Big Data*, Boston, 11-14 December 2017, 1882-1891. https://doi.org/10.1109/BigData.2017.8258133

[15] Pelech, T. and Duda, J.T. (2006) Event Detection in Financial Time Series by Immune-Based Approach. In: *Intelligent Information Processing and Web Mining*, Springer, Berlin, Heidelberg, 365-369. https://doi.org/10.1007/3-540-33521-8_38

[16] Teymourian, K., Rohde, M. and Paschke, A. (2012) Knowledge-Based Processing of Complex Stock Market Events. In: *Proceedings of the* 15*th International Conference on Extending Database Technology*, ACM, New York, 594-597. https://doi.org/10.1145/2247596.2247674

[17] Xie, C., Chen, Z. and Yu, X. (2006) Sequence Outlier Detection Based on Chaos Theory and Its Application on Stock Market. In: *International Conference on Fuzzy Systems and Knowledge Discovery*, Springer, Berlin, Heidelberg, 1221-1228. https://doi.org/10.1007/11881599_153

[18] Saidane, M. and Lavergne, C. (2008) A New Online Method for Event Detection and Tracking: Empirical Evidence from the French Stock Market. *American Journal of Finance and Accounting*, **1**, 20-51. https://doi.org/10.1504/AJFA.2008.019877

[19] Romero Meza, R., Bonilla, C. and Hinich, M. (2007) Nonlinear Event Detection in the Chilean Stock Market. *Applied Economics Letters*, **14**, 987-991. https://doi.org/10.1080/13504850600706024

[20] Shi, F., Zhou, Y., Kong, B., *et al.* (2017) Type Recognition of Financial Events by Incorporating Text and Time-Series Features. *International Conference on Machine Learning & Cybernetics*, Jeju, 10-13 July 2016. https://doi.org/10.1109/ICMLC.2016.7860874

[21] Xue, Y., Xu, L., Yu, J., *et al.* (2015) A Detection Method of Stock Event Source. 11*th International Conference on Semantics*, *Knowledge and Grids* (*SKG*), Beijing, 19-21 August 2015. https://doi.org/10.1109/SKG.2015.26

[22] Yuan, B., Chen, Q., Xiang, Y., *et al.* (2013) Event Detection and Recommendation Based on Heterogeneous Information. *Lecture Notes in Electrical Engineering*, **217**, 407-416. https://doi.org/10.1007/978-1-4471-4850-0_52

[23] Jiang, W., Xu, L., Yu, J., *et al.* (2018) Research and Application of Mapping Relationship Based on Learning Attention Mechanism. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Cham, 310-321. https://doi.org/10.1007/978-3-319-93034-3_25

[24] Jiang, W., Xu, L., Zhang, G., *et al.* (2017) The Discovery of the Relationship on Stock Transaction Data. In: *International Conference on Artificial Neural Networks*, Springer, Cham, 772-773.

[25] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523.